DOCUMENT RESUME

ED 430 050 TM 029 794

AUTHOR Fan, Xitao; Ping, Yin
TITLE Assessing the Effect of Model-Data Misfit on the Invariance Property of IRT Parameter Estimates.
PUB DATE 1999-04-00
NOTE 37p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Estimation (Mathematics); *Goodness of Fit; High Schools; *Item Response Theory; Models; State Programs; Testing Programs
IDENTIFIERS *Invariance; Large Scale Programs; Parameter Identification

ABSTRACT

This study empirically investigated the potential negative effect of item response theory (IRT) model-data misfit on the degree of invariance of: (1) IRT item parameter estimates (item difficulty and discrimination); and (2) IRT person ability parameter estimates. A large-scale statewide assessment program test database was used, for which the one-parameter IRT has poor model-data fit, and the three-parameter model has exceptionally good model-data fit. Three examinee sampling plans were used to investigating the effect of model-data misfit on the invariance of item parameter estimates, and two test sampling plans were used for investigating the effect of model-data misfit on the invariance property of IRT person ability parameter estimates. Overall, the results failed to confirm that model-data misfit in an IRT application is related to the invariance property of IRT item/person parameter estimates. Major limitations of the study are noted, and future directions are suggested. (Contains 1 figure, 7 tables, and 15 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Xitao Fan
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

MISFIT.v1

ED 430 050

Assessing the Effect of Model-Data Misfit on the Invariance Property of IRT Parameter Estimates

Xitao Fan

Yin Ping

Utah State University

Running Head: Invariance of IRT Parameter Estimates

Note. Please send correspondence about this paper to:

Xitao Fan, Ph.D.
Education Building Room 487
Department of Psychology
Utah State University
Logan, Utah 84322-2810

Phone: (435)797-1451
Fax: (435)797-1448
E-Mail: fafan@cc.usu.edu

TM029794

Paper presented at the 1999 Annual Meeting of the American Educational Research Association, April 19-23, Montreal, Canada (Session # 38.05).

BEST COPY AVAILABLE

Abstract

This study empirically investigated the potential negative effect of IRT model-data misfit on the degree of invariance of (1) IRT item parameter estimates (item difficulty and discrimination), and (2) IRT person ability parameter estimates. A large-scale state-wide assessment program test database was used, for which one-parameter IRT model has poor model-data fit, and three-parameter model has exceptionally good model-data fit. Three examinee sampling plans were used for investigating the effect of model-data misfit on the invariance of item parameter estimates, and two test item sampling plans were used for investigating the effect of model-data misfit on the invariance property of IRT person ability parameter estimates. Overall, the results failed to confirm that model-data misfit in an IRT application is related to the invariance property of IRT item/person parameter estimates. Major limitations of the study are noted, and future directions are suggested. *Index terms: item response theory, model-data fit, invariance of item parameter estimates, invariance of person ability estimates.*

The past few decades have witnessed an exponential growth in the application of item response theory (IRT) in a variety of measurement situations (Crocker & Algina, 1986; McKinley & Mills, 1989; Fan, 1998). As a result, a voluminous body of research literature related to IRT methods and their applications has been accumulated. A review of this voluminous body of IRT literature, however, reveals that there is a lack of empirical research about the potential consequences of IRT model-data misfit on the invariance property of IRT parameter estimates. In other words, although it is theoretically assumed that, in any application of IRT, the model-data fit is important, it is unclear what effect model-data misfit would have on the invariance property of IRT parameter estimates. This study was designed to explore empirically the potential effect IRT model-data misfit has on the invariance property of IRT item and person parameter estimates. Specifically, we are interested in the question: to what extent will the invariance property of IRT item and person parameter estimates be threatened by of model-data misfit in an IRT application? Before we describe the details of this study, we briefly review some relevant issues concerning IRT models in general.

IRT Models

Although classical test theory (CTT) has served the measurement community for most of this century, it has been recognized that CTT has some major weaknesses. The most important limitation of CTT can be described as a situation of circular dependency: (1) the person statistic (i.e., observed score) is (item) sample-dependent, and (2) the item statistics (i.e., item difficulty and item discrimination) are (examinee) sample-dependent. This circular dependency poses both theoretical and practical difficulties in CTT's application in many measurement situations, such as test equating and computerized adaptive testing.

Item response theory (IRT) was developed primarily in response to the problem of circular dependency for both item and person parameters of CTT. IRT framework encompasses a family of models, and the applicability of each model in a specific situation depends on the nature of the test items and the viability of different theoretical assumptions about the test items. For test items that are dichotomously scored, three IRT models are popular, and they are known as three-, two-, and one-parameter IRT models respectively. Although one-parameter model is the simplest of the three, it would be better to start from the most complex, the three-parameter, IRT model, for reasons that will be obvious momentarily. IRT three-parameter model takes the form:

$$P_i(\theta) = c_i + (1-c_i)\frac{e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}} \qquad (1)$$

where, c_i is the guessing factor, a_i is the item discrimination parameter commonly known as item slope, b_i is the item difficulty parameter commonly known as the item location parameter, D is an arbitrary constant (normally $D = 1.7$), and θ is the ability level of a particular examinee. The item location parameter is on the same scale of ability θ, and it takes the value of θ at which an examinee with the ability level θ has 50/50 chance to score the item correctly. The item discrimination parameter is the slope of the tangent line of the item characteristic curve at the point of the location parameter.

When the guessing factor is assumed or constrained to be zero ($c_i = 0$), the three parameter model is reduced to the two-parameter model for which only item location and item slope parameters need to be estimated:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (2)$$

If another restriction is imposed that stipulates that all items have equal and fixed discrimination, then a_i becomes a, a constant rather than a variable, and as such, it does not require any estimation, and the IRT model is further reduced to:

$$P_i(\theta) = \frac{e^{Da(\theta - b_i)}}{1 + e^{Da(\theta - b_i)}} \quad (3)$$

So for one-parameter IRT model, constraints have been imposed on two item parameters, and item difficulty remains the only item parameter that requires empirical estimation. This one-parameter model is also widely known as the Rasch model. It is clear from the previous discussion that the three-parameter model is the most general model, and the other two IRT models (two- and one-parameter models) can be considered as models nested[1] or subsumed under the three-parameter model.

Invariance Property of IRT Parameters

Theoretically, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person parameters. As a result, IRT models produce item parameters that are independent of examinee populations, and person parameters that are independent of the particular set of items administered. Thus, if the IRT assumptions are met and correct IRT model is applied to test data, the item parameter estimates obtained from one group of examinees should

[1] Statistically, Model B is nested under Model A if Model B can be obtained by imposing certain constraints on Model A.

remain stable across other groups, and the person ability parameter estimate θ obtained from one set of test items should also remain stable across different sets of test items.

This invariance property of item and person statistics of IRT has been illustrated theoretically (e.g., Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991), and has been widely recognized and accepted in the measurement community as the major advantage of IRT models over CTT. The invariance property of IRT model parameters makes it both theoretically and practically much easier to solve some measurement problems that are difficult to handle within the CTT framework, such as those encountered in test equating and computerized adaptive testing (Hambleton, et al., 1991). As the cornerstone of IRT, the importance of the invariance property of IRT model parameters can never be overstated, because, without this important property, the complexity of IRT models can hardly be justified on either theoretical or practical grounds.

It has been emphasized, however, that the invariance property only holds when the applied IRT model fits the data (e.g., Hambleton, 1993; Hambelton, et al., 1991). Given the importance of the invariance property of IRT models, and the emphasis placed on the good model-data fit in IRT applications (Reise, 1990), it is logical to expect that misfit between an IRT model and empirical data may potentially threaten the invariance property for IRT model parameters. As discussed by Hambleton (1993), "The advantages claimed for item response models can be realized only when the fit between the model and the test data set of interest is satisfactory. A poorly fitting model cannot yield invariant item- and ability-parameter estimates." (p. 172). Put differently, it is reasonable to expect that IRT item/person parameter estimates will exhibit higher degree of invariance when the IRT model fits the data well, and lowere degree of invariance when the IRT model fits the data poorly. Our review of the voluminous IRT literature reveals that there

appears to be a research vacuum about this issue.

Checking the Invariance Property of IRT Parameter Estimates

Although the invariance property of item/person parameterestimates of IRT models has been illustrated theoretically (e.g., Hambleton & Swaminathan, 1985; Rudner, 1983), this issue has received relatively little empirical exploration. Miller and Linn (1988), using an extant large data set, reported the results of a study which examined the variations of item characteristic functions in the context of instructional coverage variations. They reported relatively large differences in item curve responses, suggesting lack of invariance of IRT item parameters for the data they examined. Lack of invariance was also reported for IRT-based item difficulty estimates by Cook, Eignor, and Taft (1988). Other than the very limited number of studies, there appears to be an obvious lack of systematic investigation about the issues concerning the invariance property for IRT item/person parameter estimates.

Theoretically, under the condition of good fit between IRT model and test data, IRT model item parameters and person ability parameter are invariant. That is to say, (1) person ability scores based on two different sets of items are on the same scale, and (2) item parameters (e.g., item difficulty) based on two different examinee populations are also already on the same scale. In practice, however, because it is not always certain if the correct IRT model has been applied to the test data of interest, the degree of invariance of IRT model parameter estimates in an application should be considered as an issue that should be checked in research practice (Hambleton, 1993, p. 175).

From the literature, it is not entirely clear what analytic approach is the most appropriate for checking the invariance property of IRT model parameters in research practice. A common approach discussed in the literature suggests that the degree of linear relationship between

estimates of the same IRT model parameter between two examinee groups (for item parameters) or between two sets of test items (for person ability parameter) should be examined for this purpose. Hambleton (1993, pp. 175-178) and Hambleton et al., (1991, Chapter 4) provided details and examples about this approach. This approach is consistent with the general consensus that IRT "...parameters are invariant within a linear transformation" (Weiss & Yoes, 1991, p. 89).

Research Objectives

The major criticism for CTT is its inability to produce item/person parameter estimates that would be invariant across examinee/item populations. This criticism has been the major impetus for the development of IRT models, and for the exponential growth of IRT research and applications in recent decades. Given this background, it is somewhat surprising that empirical studies examining the invariance characteristics of IRT model parameters are few and scarce. It is also somewhat surprising that, in our literature review, we failed to locate any empirical study that examined the effect of model-data misfit on the invariance property of IRT model parameter estimates. Hambleton et al., (1991) commented, "In many IRT applications reported in the literature, model-data fit and the consequences of misfit have not been investigated adequately." (p. 53). It is also this lack of empirical investigation that has prompted some researchers to state that item response modeling has been too concerned with the mathematical elaboration at the expense of empirical exploration (Goldstein & Wood, 1989). Studies that investigated the consequences of model-data misfit typically focused on the potential bias such model-data misfit might cause to IRT parameter estimates (e.g., Meijer & Nering, 1997).

The present study was designed to explore the issues about the consequences of IRT model-data misfit. More specifically, we focused on the invariance property of IRT parameters, and attempted to address three questions:

1. What negative effect does IRT model-data misfit have on the invariance property of IRT item difficulty parameter estimates?
2. What negative effect does IRT model-data misfit have on the invariance property of IRT item discrimination parameter estimates?
3. What negative effect does IRT model-data misfit have on the invariance property of IRT person ability parameter estimates?

Methods

Data Source

The data used in this study are from the Texas Assessment of Academic Skills (TAAS) tests administered in October 1992 and taken by 11th Grade students at the time. Designed for assessing the mastery of school instructional objectives, TAAS was a state-mandated criterion-referenced test battery consisting of Reading, Math and Writing tests. Used in this study were the Reading (48 items) and the Math (60 items) tests that consisted of multiple-choice items scored dichotomously as either correct or incorrect. Unattempted items were scored as incorrect responses. The examinee pool for the database has over 193,000 subjects. Table 1 presents the demographic information of the examinee pool of this database.

Insert Table 1 about here

TAAS was designed to be a test battery for assessing minimum-competency of students in Texas public schools. As is typically the case for mastery tests, TAAS test items were primarily curriculum content-based, and test score distributions were negatively skewed (skewness= for Reading, and skewness= for Math tests), indicating some ceiling effect of the score distributions.

Examinee Sampling

To examine the degree of invariance of IRT item parameter estimates (item difficulty and item discrimination), three sampling plans were implemented for Math and Reading test data so that the behaviors of IRT item parameter estimates could be examined under different examinee sample conditions. The three sampling plans generated samples that were progressively more dissimilar from each other, and this sampling strategy allowed the examination of the behaviors of IRT item parameter estimates across progressively less comparable examinee samples. All examinee samples in this study had sample size of 1,000, which is generally considered to be sufficiently large for estimating IRT parameters in the three-parameter IRT model.

Random samples. Random samples of examinees, each consisting of 1,000 examinees, were drawn from the entire subject pool. Twenty such samples were drawn for Math test data, and 20 for Reading, making the total number of random subject samples to be 40. Because these were random samples from the same population, they should be comparable with each other within the limits of statistical sampling error.

Gender group samples. Samples of female students, and those of male students, were randomly drawn separately for Math and Reading test data. Twenty female samples and twenty male samples were drawn for each test, making the total number of gender samples to be 80. Because the female and male samples were drawn from different populations as defined by the demographic variable gender, theoretically, there should be more dissimilarity between a female sample and a male sample than there is between two random samples described in the section "Random samples" above. Table 2 presents the performance statistics for the female and male groups. It is seen that the female and male groups had comparable performance for Reading, while there is a slight difference for Math at the test level.

Insert Table 2 about here

High- and low-ability group samples. This sampling plan generated samples which were different in terms of performance on the tests. High-ability group was defined as those whose scores fell within the 15th to 100th percentile range (15%ile - 100%ile) on Math or Reading test. Low-ability group was defined as those whose scores fell within the 0th to 85th percentile range (0%ile - 85%ile) on Math or Reading test. Twenty samples were randomly drawn from each of the two group, and separately for each test, making the total number of high-ability and low-ability samples to be 80. Because these two groups were defined in terms of test performance, not in terms of a demographic variable as in gender group sampling, it is logical to expect that there should be more dissimilarity between a high-ability sample and a low-ability sample than there is between a female and a male sample pair.

Degree of invariance of IRT item parameter estimates. The issue about the degree of invariance of IRT item parameter estimates is a crucial one for this study. For assessing the degree of invariance of IRT item parameter estimates, we followed the discussion by Weiss & Yoes (1991) that "In practice, ... the item parameters are invariant within a linear transformation" (p. 89), and we adopted the practical approaches discussed by Hambleton (1993) and Hambleton et al., (1991), and conducted correlation analyses for item parameter estimates derived from two different examinee samples. The three examinee sampling plans discussed previously allowed us to assess the degree of invariance of IRT item parameter estimates across progressively dissimilar samples: between two random samples of the same population, between female and male samples, and between high- and low-ability samples.

Test Item Sampling

To examine the invariance property of IRT person ability parameter (θ), two item sampling plans were used. The two sampling plans generated test item samples which were progressively more dissimilar, thus allowing the examination of the degree of invariance of IRT person ability estimates across progressively less comparable test item samples.

Randomly splitting the test items into two tests. The 60 Math test items were randomly split into two 30 item pools, and the two 30-item pools were used as two parallel math. The person ability estimates were obtained from each of these two math tests. Because the two math tests were constructed based on two random samples of test items, the two tests were comparable with each other within the limits of sampling error. As a result, the person ability parameter estimates from these two tests should also be considered comparable within the limits of sampling error. The same random splitting procedure was applied to the TAAS Reading test items (total of 48), and two parallel reading tests were constructed, with each test consisting of 24 randomly sampled reading items.

Twenty independent examinee samples with sample size of 1,000 for each were drawn from the data base. For each examinee sample, IRT person ability estimates under one-, two, or three-parameter IRT models were obtained for each examinee based on the two parallel tests. Degree of invariance of IRT person ability estimates were assessed through the empirical relationship between person ability parameter estimates obtained from the two parallel tests under each of the three IRT models.

Splitting the test items into easy and difficult tests. For this non-random item sampling plan, based on a subset of 20,000 examinees, the 60 TAAS Math items were ranked from the most difficult to the easiest based on the p-values of the items. Once ranked on item difficulty,

the 60 items were divided into four quartiles, with the first quartile being the most difficult one-fourth of items, and the last quartile being the easiest one-fourth of items. The first and the third quartile of the items were assigned to the "difficult test", and the second and the fourth quartiles of the items were assigned to the "easy test". The same non-random item splitting procedure was applied to the 48 TAAS Reading items to create a "difficult" reading test and an "easy" reading test. Because the "difficult" and "easy" tests were constructed using the non-random sampling procedure to divide the test items into "difficult" and "easy" tests, the two tests ("difficult" vs. "easy" tests, for both math and reading items) were systematically more dissimilar than the two tests created based on random item sampling procedure described in the previous section. Table 3 presents the descriptive information about the items for the "difficult" and "easy" tests for math and reading.

Insert Table 3 about here

Twenty independent examinee samples with sample size of 1,000 for each were drawn from the data base. Fro each examinee sample, estimates of IRT person ability parameter were then obtained for each examinee based on both the "difficult" and the "easy" tests, utilizing one-, two, and three-parameter IRT models. Finally, degree of invariance of IRT person ability estimates are assessed through the empirical relationship between ability parameter estimates (θ) based on the "difficult" and "easy" tests under one of the three IRT models.

Results and Discussions

The results of the study are presented in the order of the three research questions presented earlier. Whenever appropriate, relevant interpretation and discussion about the

meaning and implications of the results are presented together with the results. But before the results related to the research questions are presented, the question of IRT model fit should be addressed.

IRT Model Fit Assessment

In any application of IRT model, it is always important to assess the extent to which the IRT model assumptions are valid for the given data, and the degree of model-data fit between the IRT model used and the given test data. The violation of IRT model assumptions, and the misfit between the IRT model and the testing data, may lead to erroneous unstable, or biased IRT model parameter estimates. In the present study, the assessment of IRT model fit was conducted on a simple random sample of 6,000 examinees for TAAS Math and Reading tests (equal sample size for the two tests, but different samples). The large sample size used should have provided stable and trustworthy results about the model assumption and model fit.

Unidimensionality is an important assumption common for all three IRT models used in this study. Typically, the validity of this assumption for the given data is empirically assessed by investigating if a dominant factor exists among all the items of the test (Hambleton, et al., 1991). A common and reasonable procedure used in this situation is exploratory factor analysis, and its results related to the eigenvalues of all the factors. Figure 1(a) presents the "scree plot" for the first seven (the largest seven) eigenvalues for the 48 test items on TAAS Reading test. Figure 1(b) presents the "scree plot" for the first seven (the largest seven) eigenvalues for the 60 test items on TAAS Math test. It is obvious that a very dominant factor exists in both situations. Other than the first dominant factor, all the other factors can be concluded as representing the "scree", that is, unimportant factors. Based on these results, it is reasonable to conclude that the unidimensionality assumption for the IRT models holds for the data used in this study.

Insert Figure 1 about here

The model-data fit was assessed by checking if individual test items misfit a given IRT model. The likelihood-ratio χ^2 test in BILOG V3.07 (Mislevy & Bock, 1990) that assesses the discrepancy between the expected response pattern and the actual response pattern of the subjects on a particular item is conducted for each item. Table 4 summarizes the number of items identified as misfitting the given IRT model at the α=.01 level.

Insert Table 4 about here

It should be pointed out that, given the examinee sample size of 6,000 used in the analyses for assessing the IRT model fit, the statistical test for identifying misfitting items has considerable statistical power. It is likely that a relatively small difference between the expected and the empirical response pattern would have flagged an item as a misfitting item for a given IRT model. For both TAAS Math and Reading test items used in this study, only one item was identified as misfitting the three-parameter IRT model, indicating that the three parameter IRT model fits the test data exceptionally well. The model-data fit for the one-parameter model, however, is obviously much worse, with about close to 50% of the items being identified as misfitting the IRT one-parameter model for both TAAS Math and Reading tests. The model-data fit for the two-parameter model is slightly worse than that for the three parameter model, but substantially better than that for the one-parameter model. Given the information in Table 4 about model-data fit for the data used in this study, if model-data misfit threatens the invariance property of IRT parameter

estimates, it would be expected that the degree of invariance would be the highest for the three-parameter IRT model parameter estimates because of its best model-data fit, and the lowest for the one-parameter IRT model parameter estimates because of its worst model-data fit.

Research Question #1

Table 5 presents the results related to the first research question, "What negative effect does IRT model-data misfit have on the invariance property of IRT item difficulty parameter estimates?". The average correlation coefficients in this table are average correlations between IRT item difficulty parameter estimates derived from two different examinee samples. For example, for one-parameter IRT model and for Between Female-Male Samples, the entry for Math test is .947. This is the average of the correlations between IRT item difficulty estimates from a female sample and those obtained from a male sample. One hundred such female-male sample pairs were formed and IRT item difficulty estimates correlated between the female and male samples within each pair. These one hundred correlation coefficients were then averaged to be .947 through Fisher z transformations. Other entries in the table were obtained in the same fashion. It is important to note that invariance property of item parameters is investigated by administering the same items to different examinee samples, and then examining the relationship between item parameter estimates obtained from different examinee samples.

Insert Table 5 about here

It is observed that, under all three sampling plans, for both TAAS Math and Reading tests, the average between-sample correlations for one-parameter IRT item difficulty parameter estimates are all slightly higher than those for two- and three-parameter IRT item difficulty

parameter estimates. These results appear to suggest that, for the data given, one-parameter IRT item difficulty estimates are slightly more "invariant" across samples than the two- and three-parameter IRT model item difficulty estimates. Considering that invariance only holds when the fit of the model to the data is good (Hambleton, et al., 1991, p. 23), do these results imply that the one-parameter model fits the data slightly better than the two- and three-parameter models? Previous results of model fit assessment (see Table 4), however, indicate that the reverse is probably true. Also, from the perspective of statistical modeling, it is also somewhat unlikely that one-parameter IRT model fits the data better than two- or three-parameter models, because one-parameter model can be considered as a submodel nested under the two- or three-parameter models. Theoretically, a model higher in a model hierarchy tends to provide better fit than a model nested under it, because the lower model has more constraints. A constrained parameter will tend to increase the misfit of the model, and the question is usually "how much?". If the misfit caused by the constrained parameter is minimal relative to the gain in model parsimony, the simpler and more restrictive model will be preferred.

In general, the degree of invariance of IRT item difficulty parameter estimates is quite high (average correlation coefficients around .95 except one or two cases) for all three IRT models under the three different examinee sampling plans. But contrary to our expectation that model-data misfit may threaten the invariance property of IRT item difficulty parameter estimates, the results in Table 5 do not reveal that model-data misfit for the one-parameter IRT model has caused any observable negative effect on the degree of invariance for the IRT item difficulty parameter estimates. If anything, the reverse is observed: the item difficulty parameter estimates for the one-parameter IRT model appear to be slightly more invariant across examinee samples than the better fitting two- and three-parameter IRT model item difficulty estimates.

Research Question #2

The second research question asks, "What negative effect does IRT model-data misfit have on the invariance property of IRT item discrimination parameter estimates?". Table 6 presents the results of correlation analyses for item discrimination parameter estimates for two- and three-parameter IRT models. As explained before, because IRT one-parameter model (Rasch model) assumes fixed item discrimination parameter for all items, no correlations could be computed for one-parameter IRT model between examinee samples, hence the N/As (Not Applicable) under IRT 1P in the table. Again, each table entry is the average of 100 correlation coefficients obtained from 100 sample pairs between IRT item slopes across two examinee samples.

Insert Table 6 about here

It is noted that IRT item discrimination parameter estimates are less invariant across examinee samples than IRT item difficulty parameter estimates presented in Table 5, with the average correlation coefficients being lower, and in a few cases, substantially lower, than .90. The results here bare some resemblance to those reported by Sireci (1991) that IRT had stable item difficulty parameter estimates, but could not successfully provide stable item discrimination parameters, although Sireci's results were based much smaller sample sizes. In most cases, the average between-sample correlations of IRT item discrimination parameter estimates are moderately high (high .80s to low .90s), indicating reasonable invariance across examinee samples. Because it is not clear at all from the literature what criteria are available for judging the degree of invariance of IRT model parameter estimates, our use of "reasonable invariance" here is

inherently subjective. But the invariance of IRT item discrimination indices decreases with the increasing dissimilarity between examinee samples. In other words, the IRT item discrimination parameter estimates are most invariant across random samples, less invariant across female-male samples, and least invariant across high-low ability samples. The difference in the degree of invariance between two- and three-parameter IRT model item discrimination parameter estimates is not large, and neither is the difference consistent in the direction.

For the last condition (Reading Test, Between High-Low Ability Samples), the IRT two-parameter model item slopes maintained moderate degree of invariance (r=.635), but the degree of invariance for three-parameter IRT model item slopes was obviously quite low (r=.321). This observation is somewhat puzzling. As discussed above, theoretically, if parameters could be adequately estimated for the given sample size, a higher order (less restrictive) model tends to provide better fit than a lower order (more restrictive) model, although such better fit may come at the expense of model parsimony. If a better fit is obtained, higher degree of "invariance" of item parameters would be expected (Hambleton, et al., 1991). The observation that, in this situation, the two-parameter IRT model had moderately invariant item discrimination indices and the three-parameter IRT model item discrimination indices showed little invariance property for the same data is contrary to both intuition and theoretical expectation. This result may indicate that the estimation for the IRT model item discrimination parameter for the given data might be somewhat unstable.

Back to the research question, "What negative effect does IRT model-data misfit have on the invariance property of IRT item discrimination parameter estimates?", the answer here is far from clear. As presented in Table 4, one-parameter model has the worst most model-data fit for the test data. But for one-parameter model, item discrimination is fixed, thus the issue of

"invariance" is irrelevant. Between the two- and three-parameter IRT models, however, the difference in model-data fit is small (for Math test, 2 vs. 1 items identified as misfitting items for two- and three-parameter IRT models respectively; for Reading test, 7 vs. 1 items identified as misfitting items for two- and three-parameter IRT models, respectively). In addition, the difference in the degree of invariance between the two- and three-parameter model item discrimination estimates is not consistent either. If the difference in model-data misfit between two- and three-parameter IRT models were more obvious, and the difference in the degree of invariance between two- and three-parameter model item discrimination estimates were more consistent in direction, more definitive answer to this research question might be possible.

Research Question #3

The third research question asks, "What negative effect does IRT model-data misfit have on the invariance property of IRT person ability parameter estimates?". Table 7 presents the results related to this question. Because of the substantial model-data misfit of one-parameter IRT model, and the excellent model-data fit for the three-parameter model, if model-data misfit threatens the invariance property of IRT person ability parameter estimates, we would expect lower degree of invariance for one-parameter IRT person ability estimates than that for three-parameter IRT person ability estimates, with the two-parameter model somewhere in between.

Insert Table 7 about here

In general, the correlation analysis findings presented in Table 7 show very small differences among the one-, two-, and three-parameter IRT model results. The correlation coefficients between person ability parameter estimates derived from two different tests are very

similar across the three IRT models, and for both Math and Reading tests, indicating that there is similar degree of invariance for IRT person ability parameter estimates for the three IRT models. For example, for Math, the average correlation between person ability estimates obtained from "difficult" and "easy" tests is .847 for one-parameter IRT model, .854 for two-parameter IRT model, and .858 for three-parameter IRT model.

Although the difference in the degree of invariance for person ability estimates among the three IRT models is very small, a close look reveals that there is a tendency that three-parameter model IRT person ability estimates show slightly higher degree of invariance than those of two-parameter IRT model, which in turn, show slightly higher degree of invariance than those of one-parameter IRT model. This tendency appears to exist for both item sampling plans (between random-split tests, and between "difficult" and "easy" tests), and for both Math and Reading tests. This tendency is consistent with the expectation that one-parameter model person ability estimates should have lower degree of invariance than three-parameter model person ability estimates for the data used, because one-parameter IRT model has the worst fit for the data, and three-parameter model has the best fit (see results in Table 4).

It is also noted that Reading test items have lower degree of invariance (correlation coefficients between two tests in the upper .70 range) for the person ability estimates than Math test (correlation coefficients between two tests in the middle and upper .80 range). Reading tests contained easier test items than the math tests (see Table 3 for average item p-values for Math and Reading tests). When test items are either too easy (as in this case) or too difficult relative to the examinee ability, item information is typically low (Hambleton, et. al., 1991, Chapter 6), and person ability estimates may be less accurate. Consequently, person ability estimates may show lower degree of invariance across different test items.

Summary and Conclusions

This study empirically examined the issue about whether IRT model-data misfit threatens the invariance property of IRT model parameter estimates. The study focused on the potential negative effect of IRT model-data misfit on the degree of invariance of (1) IRT item parameter estimates (both item difficulty and item discrimination parameter estimates), and (2) IRT person ability parameter estimates. A large-scale test database from a state-wide assessment program was used as data source for the investigation. The test item pool had two tests with 60 and 48 dichotomously scored items in each, and the examinee pool had more than 193,000 examinees.

Preliminary analyses show that one-parameter IRT model has the worst model-data fit, and three-parameter model has the best model-data fit for the data used in this study. For investigating the potential negative effect of IRT model-data misfit on the invariance of item parameter estimates, examinee sample pairs (n=1,000 for each sample) were drawn from the examinee pool under three sampling plans (random, male-female, and high-low ability examinee samples), producing progressively more dissimilar examinee sample-pairs to facilitate the assessment of the degree of invariance of IRT item parameter estimates under one-, two-, and three-parameter IRT models. Item parameter estimates were obtained, and they were correlated across the two examinee samples within each sample pair under each of the three IRT models (one-, two-, and three-parameter models).

For investigating the potential negative effect of IRT model-data misfit on the invariance property of IRT person ability parameter estimates, two sampling plans for test items (random test item samples, and difficult-easy test item samples) were used. Two tests were constructed under each sampling plan so that an examinee's ability could be estimated from both tests. Under each of the three IRT models (one-, two-, and three-parameter models), person ability estimates were

obtained from two tests (two randomly split tests, or "difficult" vs. "easy" tests), and were correlated.

The major findings are as follows:

(1) For the data used in this study, IRT item difficulty indices have exhibited very high degree of invariance across samples, even across samples which were quite different (samples from high- and low-ability groups). No negative effect of model-data misfit on the invariance property of IRT item difficulty parameter estimates was observed in the results. Contrary to our expectations, one-parameter IRT model, which has the worst model-data fit, has exhibited a tendency of having slightly more invariant item difficulty parameter estimates than the better fitting two- and three-parameter IRT models.

(2) For the data used in this study, IRT item discrimination parameter estimates are generally less invariant than their item difficulty counterparts. The degree of invariance of item discrimination parameter estimates decreases steadily as examinee sample pairs became more dissimilar, implying that IRT item discrimination parameter estimates may not maintain a high degree of invariance across populations that are sufficiently different. Because the issue of invariance of item discrimination parameter is irrelevant for one-parameter IRT model, only two- and three-parameter model could be used for comparing the degree of invariance of item discrimination parameter estimates. The results provided no interpretable findings or implications about the potential negative effect of model-data misfit on the invariance property of IRT item discrimination parameter estimates.

(3) For the data used in this study, IRT person ability parameter estimates showed very small differences in the degrees of invariance across the three IRT models. A close look at the results, however, reveals that, consistent with the theoretical expectations, three-parameter IRT

model person ability estimates showed slightly higher degree of invariance than those of two-parameter IRT model estimates, which in turn, showed slightly higher degree of invariance than one-parameter IRT model person ability estimates.

Overall, the results of this study are inconclusive about the potential negative effect of model-data misfit on the invariance property of IRT item/person parameter estimates. For IRT person ability parameter estimates, there appears to be a slight tendency that model-data misfit might reduce the degree of invariance of IRT parameter estimates. On the other hand, for IRT item difficulty parameter estimates, better fitting IRT models (two- and three-parameter models) did not produce more invariant item difficulty parameter estimates than one-parameter IRT model with the worst model-data fit.

As discussed previously in the article, the invariance property of IRT item/person parameter estimates is an important issue for IRT models. Our review of the literature, however, indicates that issues related to the invariance property of IRT item/person parameters have not been adequately investigated. Consequently, not much appears to be known about the robustness of the invariance property of IRT item/person parameters when model-data fit is poor. Although the findings in this study are inconclusive, they do raise the questions about how and to what extent model-data misfit is related to the invariance property of IRT item/person parameters.

Limitations

This study, like many others, has its share of limitations that may potentially undermine the validity of its findings. First of all, the characteristics of the test items used in the study may be somewhat unique. As discussed in the Methods section and indicated in Table 3, the test items tend to be easy for the examinees, and the score distributions exhibited ceiling effects, as is generally the case for minimum-competency tests or other criterion-referenced mastery tests.

Although it is unclear what systematic impact this data characteristic may have on the results, it would be desirable in future studies to use data from norm-referenced testing that has items varying more in item difficulty, and likely varying more in item discrimination also.

The second shortcoming of this investigation is the limited item pool used in the study and the uncontrolled item characteristics. Although the examinee pool is adequate in the sense that a variety of different examinee samples can be drawn from it, the same cannot be said about the item pool. Ideally, the test item pool should be larger and more diverse in terms of item characteristics so that the behaviors of IRT item/person parameter estimates can be studied under different conditions of item characteristics. Future studies may benefit from using several different testing databases. More importantly, Monte Carlo simulation studies that give the researcher complete control over different aspects of item characteristics have the potential to provide more definitive answers to the questions raised in this study.

References

Cook, L. L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. Journal of Educational Measurement, 25, 31-45.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart & Winston.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. Educational and Psychological Measurement, 58, 357-381.

Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Hambleton, R. K. (1993). Principles and selected applications of item response theory. In R. Linn (Ed.), Educational measurement (3rd. ed.), pp. 147-200. Phoenix, AZ: The Orxy Press.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Meijer, R. R., & Nering, M. L. (1997). Trait level estimation for nonfittting response vectors. Applied Psychological Measurement, 21, 321-336.

McKinley, R. & Mills, C. (1989). Item response theory: Advances in achievement and attitude measurement. In B. Thompson (Ed.), Advances in social science methodology (Vol. 1, pp. 71-135). Greenwich, CT: JAI.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. Journal of Educational Measurement, 25, 205-219.

Mislevy, R. J., & Bock, R. D. (1990). BILOG3: Item analysis and test scoring with binary logistic models. Chicago: Scientific Software International.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. Applied Psychological Methods, 14(2), 127-137.

Rudner, L. M. (1983). A closer look at latent trait parameter invariance. Educational and Psychological Measurement, 43, 951-955.

Sireci, S. G. (1991). "Sample-independent" item parameters? An investigation of the stability of IRT item parameters estimated from small data sets. Paper presented at the Annual Conference of the Northeastern Educational Research Association (Ellenville, NY, October 24). (ERIC Document Reproduction Services No. ED 338 707).

Weiss, D. J., & Yoes, M. E. (1991). Item response theory. In R. K. Hambleton and J. N. Zeal (eds.), Advances in educational and psychological testing, pp. 69-96. Boston, MA: Kluwer Academic Publishers.

Table 1: Ethnicity and Gender Composition of the Subject Pool

Group		Frequency	%	Cumulative Frequency
Ethnicity	American Indian	526	.3	526
	Asian-American	5815	3.0	6341
	African-American	24714	12.8	31055
	Hispanic	59918	31.0	90975
	White	98166	50.8	189141
	Unknown or Not Indicated	4101	2.1	193240
Gender	Female	98240	50.8	98240
	Male	94610	49.0	192850
	Unknown or Not Indicated	390	.2	193240

Table 2 Performance Characteristics of Female and Male Groups

		Mean	STD	Q1	Median	Q3[a]
Reading	Female	37.17	7.43	33	39	43
	Male	37.48	7.48	33	39	43
Math	Female	41.81	11.02	34	43	51
	Male	43.26	11.26	36	45	53

a These are the first quartile (25th percentile), the second quartile (50th percentile, or median), and the third quartile (75th percentile), respectively.

Table 3 Item P-Values for "Difficult" and "Easy" Tests

Area	Tests	Mean	SD	Minimum	Maximum
Math	"Difficult"	.6533	.1049	.4360	.7906
	"Easy"	.7560	.1015	.6317	.9183
Reading	"Difficult"	.7100	.1276	.3908	.8653
	"Easy"	.8399	.0973	.7263	.7458

Table 4 Number of Items Identified [a] as Misfitting the IRT Models for the Two Tests

		IRT Models		
Test	# of Items	1P[b]	2P	3P
Math	60	27	2	1
Reading	48	22	7	1

a All tests for identifying misfitting items for a given IRT model were conducted at α=.01 level.

b These are the one-, two-, and three-parameter IRT models respectively.

Table 5 Invariance of IRT Item Difficulty Parameter Estimates: Average Correlations

Examinee Sampling Plan		IRT Models [a] 1P	2P	3P
Between Random Samples	Math	.988 (.002) [b]	.968 (.010)	.965 (.009)
	Reading	.991 (.002)	.966 (.012)	.969 (.009)
Between Female-Male Samples	Math	.947 (.007)	.929 (.014)	.926 (.014)
	Reading	.973 (.004)	.955 (.010)	.955 (.009)
Between High-Low Ability Samples	Math	.978 (.005)	.907 (.029)	.925 (.014)
	Reading	.979 (.003)	.862 (.029)	.877 (.025)

a For one-, two- and three-parameter IRT models, respectively.

b An average correlation coefficient was obtained through (1) transforming individual correlation coefficients to Fisher Zs, (2) averaging the Fisher Zs, and (3) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations of the original Pearson correlation coefficients is in parenthesis.

Table 6 Invariance of IRT Item Discrimination Indices: Average Correlations

Examinee Sampling Plan		IRT Models [a] 1P	2P	3P
Between Random Samples	Math	N/A	.906 (.019) [b]	.857 (.037)
	Reading	N/A	.891 (.024)	.920 (.025)
Between Female-Male Samples	Math	N/A	.877 (.023)	.837 (.029)
	Reading	N/A	.864 (.029)	.880 (.044)
Between High-Low Ability Samples	Math	N/A	.748 (.034)	.631 (.055)
	Reading	N/A	.636 (.078)	.321 (.089)

a For one-, two- and three-parameter IRT models, respectively.

b An average correlation coefficient was obtained through (1) transforming individual correlation coefficients to Fisher Zs, (2) averaging the Fisher Zs, and (3) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations of the original Pearson correlation coefficients is in parenthesis.

Table 7 Invariance of IRT Person Ability Estimates: Average Correlations between Person Ability Estimates on Two Tests

Test Item Sampling Plan		IRT Models [a] 1P	2P	3P
Between Random-Split Tests	Math	.851(.007)[b]	.860(.006)	.863(.006)
	Reading	.791(.010)	.797(.011)	.799(.011)
Between Difficult-Easy Tests	Math	.847(.008)	.854(.008)	.858(.007)
	Reading	.769(.012)	.779(.014)	.778(.012)

a For one-, two- and three-parameter IRT models, respectively.

b An average correlation coefficient was obtained through (1) transforming individual correlation coefficients to Fisher Zs, (2) averaging the Fisher Zs, and (3) transforming the average Fisher Z back to the Pearson correlation coefficient. Standard deviations of the original Pearson correlation coefficients is in parenthesis.

Figure Captions:

Figure 1 Scree Plots of the First Seven Eigenvalues of TAAS Reading (a) and Math (b) Test Items

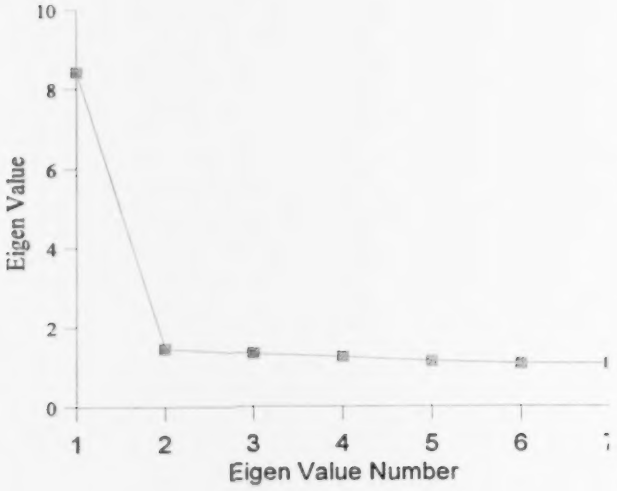

Figure 1 (a)

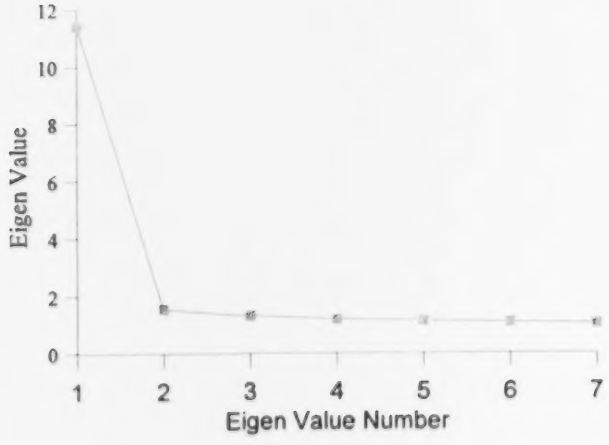

Figure 1 (b)